



How Edge Computing, Artificial Intelligence, and Generative AI are changing the future of restaurant technology



Abstract

Frequent challenges faced by the restaurant industry can be categorized under three broad level segments, including low customer retention, inefficient operations and inventory management and high labor cost.



Low customer retention: The restaurant industry is highly competitive and fragmented, with customers having a wide range of choices and preferences. The percentage of customers who are loyal to a specific restaurant brand is declining, whereas those who switch brands more than once a month is on the rise. To retain and attract customers, restaurants need to offer personalized and engaging experiences, such as customized menus, recommendations, rewards, and feedback.



Inefficient operations and inventory management: The restaurant industry faces various operational challenges, such as optimizing food quality and safety, reducing food waste and spoilage, managing supply chain and inventory, and complying with health and safety regulations. According to a report, the average food wastage in restaurants is 11% of food purchases, which amounts to significant losses annually. To improve operational efficiency and profitability, restaurants need to leverage real-time data and analytics, automate processes, and optimize resources.



High labor cost and turnover rates: The restaurant industry is one of the most labor-intensive sectors. According to a survey, 98% of operators say higher labor costs are an issue for their restaurant. Moreover, the industry suffers from a high turnover rate which impacts the quality and consistency of service and increases training and hiring costs.

This paper discusses how Edge Computing, AI, and Generative AI can help address these challenges by bringing computing power closer to where the data is generated, reducing latency, and enabling faster decision-making. It explores the advantages of Edge Computing, the use of in-restaurant cloud technology, and the benefits of using Large Language Models (LLMs) on Edge devices. The paper also discusses the technical challenges that are needed to be resolved, leveraging the methods for model quantization.



Introduction

The restaurant industry is undergoing a digital transformation, driven by the need to enhance customer experiences, optimize operations, and increase revenue. Edge Computing, AI, and Generative AI are some of the key technologies that are enabling this transformation. According to a report by Grand View Research, the global Edge Computing market size is expected to reach \$155.90 billion by 2030, growing at a compound annual growth rate (CAGR) of 36.9%.



The report also states that, “Artificial Intelligence (AI) integration into the Edge environment is projected to drive market growth. An Edge AI system is estimated to help businesses make decisions in real time in milliseconds. The need to minimize privacy concerns associated while transmitting large amounts of data, as well as latency and bandwidth issues that limit an organization’s data transmission capabilities, are factors projected to fuel market growth in the coming years.”

Generative AI: Generative AI is a branch of AI that can generate novel and realistic content, such as images, text, music, or video, based on existing data.

Edge Computing: Edge Computing is the idea of doing computing activities near where the data comes from, to reduce the delay between the data and the decisions. One of the main differences

between Edge and Cloud Computing is the location of data processing. While Cloud Computing relies on centralized servers to store and process data, Edge Computing distributes the data processing across local devices or servers that are closer to the data source. This reduces the latency, bandwidth, and privacy issues that are associated with Cloud Computing. Edge Computing can also enable more efficient and reliable AI applications that allow real-time or near real-time decision making.

Restaurants use in-restaurant Cloud technology with Edge Computing to speed up and ensure data processing and system uptime for in-store applications.

The comparison between Edge and Cloud deployments is shown in the picture below.

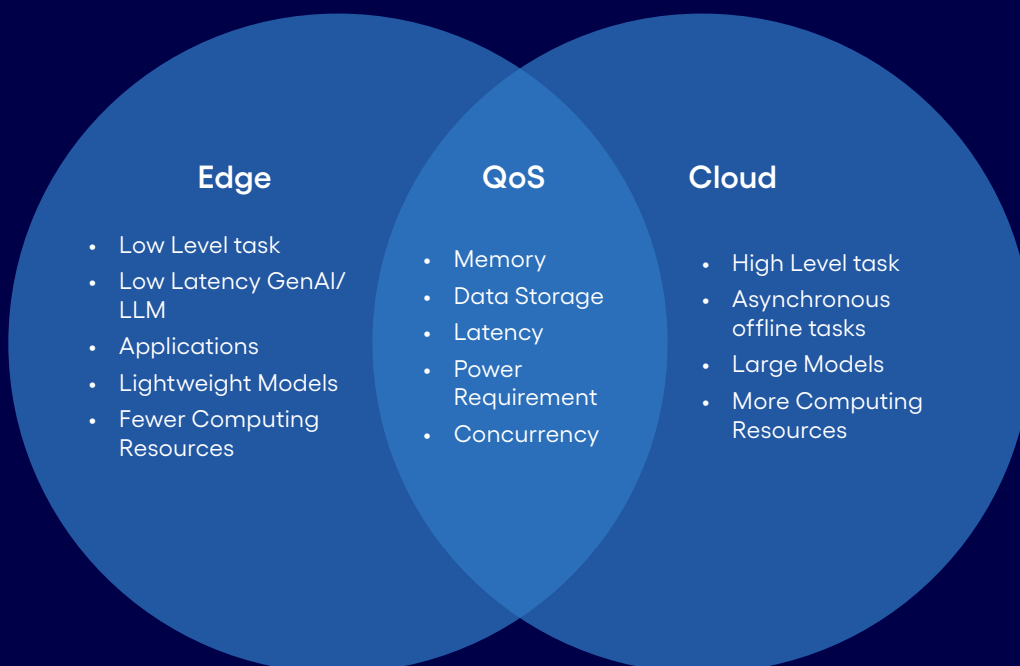


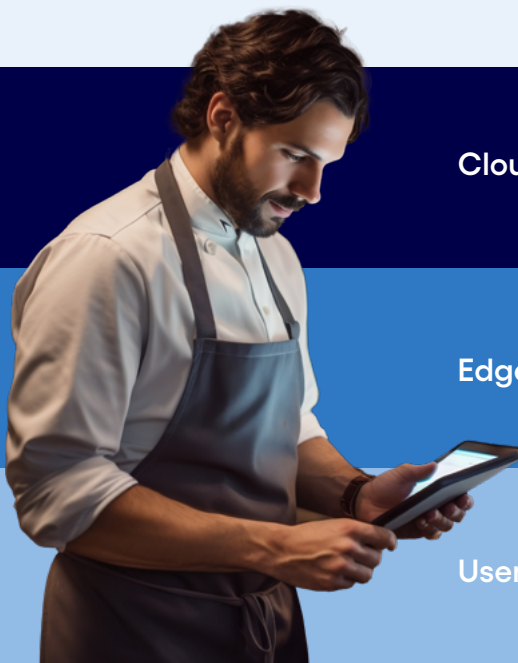
Diagram 1: Edge vs Cloud

In-restaurant cloud technology driven by Edge Computing can help restaurants process data faster and more reliably and enhance system uptime. Data is synchronized with the central cloud data store and the in-store applications can switch between the in-store cloud and the public cloud as needed. The hybrid Edge infrastructure blends the public cloud and the in-store cloud and forms the basis of the new Cloud Computing for restaurants to ensure business continuity. By analyzing data at the Edge in real-time or near real-time, businesses can train AI models and improve the performance of AI driven applications. Some of the decisions that can happen at the store level are:

- Computer vision technology, Generative AI, machine learning and deep learning frameworks for AI-driven personalization, in-restaurant housekeeping, dynamic pricing, promotion, inventory, and production optimization and various other IoT driven operations that use huge amount of data for predictive analysis.
- A small and powerful in-store device, an example of Edge Computing, brings computing power to the data required to run the restaurant operations. Also, the APIs in the public cloud that are needed for restaurant operations are copied at the in-store cloud to enable faster order management and payment processing. These APIs will keep the restaurant running even when there is no connectivity.

This kind of resilient and redundant architecture helps restaurants maintain business continuity, reliability in payment processing, which reduces financial losses and increases customer satisfaction due to faster speed and uptime.

Edge Computing is essential for the rapid development of Generative AI as it solves the problems of real-time processing, lower latency, and effective data management. As the companies deploy Generative AI solutions, they will have to deal with the issues of long-term cost, data privacy and security. Edge Computing helps to overcome these challenges and clean up the raw data before moving the data to the public cloud for more costly AI training operations.



Cloud

Edge server

User device

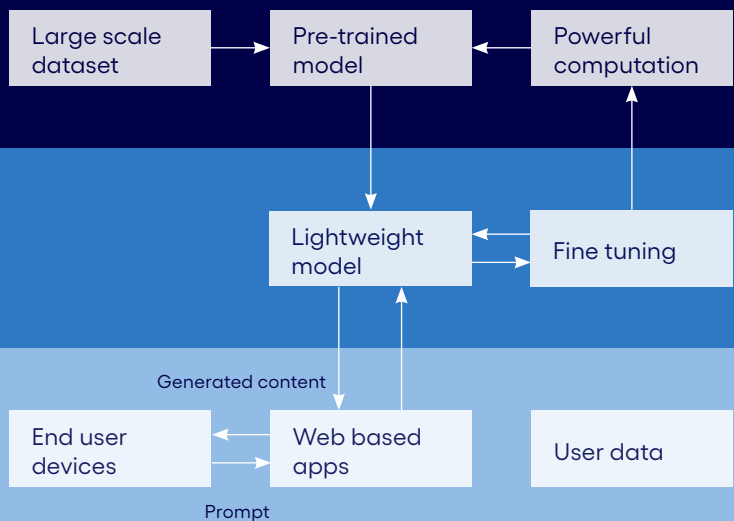


Diagram 2: Logical representation of LLM Model deployment on Edge and Cloud

How Edge Computing and GenAI can improve restaurant operations

Based on this research, the following business map depicts the modules (blue highlighted), in which the combination of Edge, GenAI and Traditional AI can have considerable impact on restaurant operations.

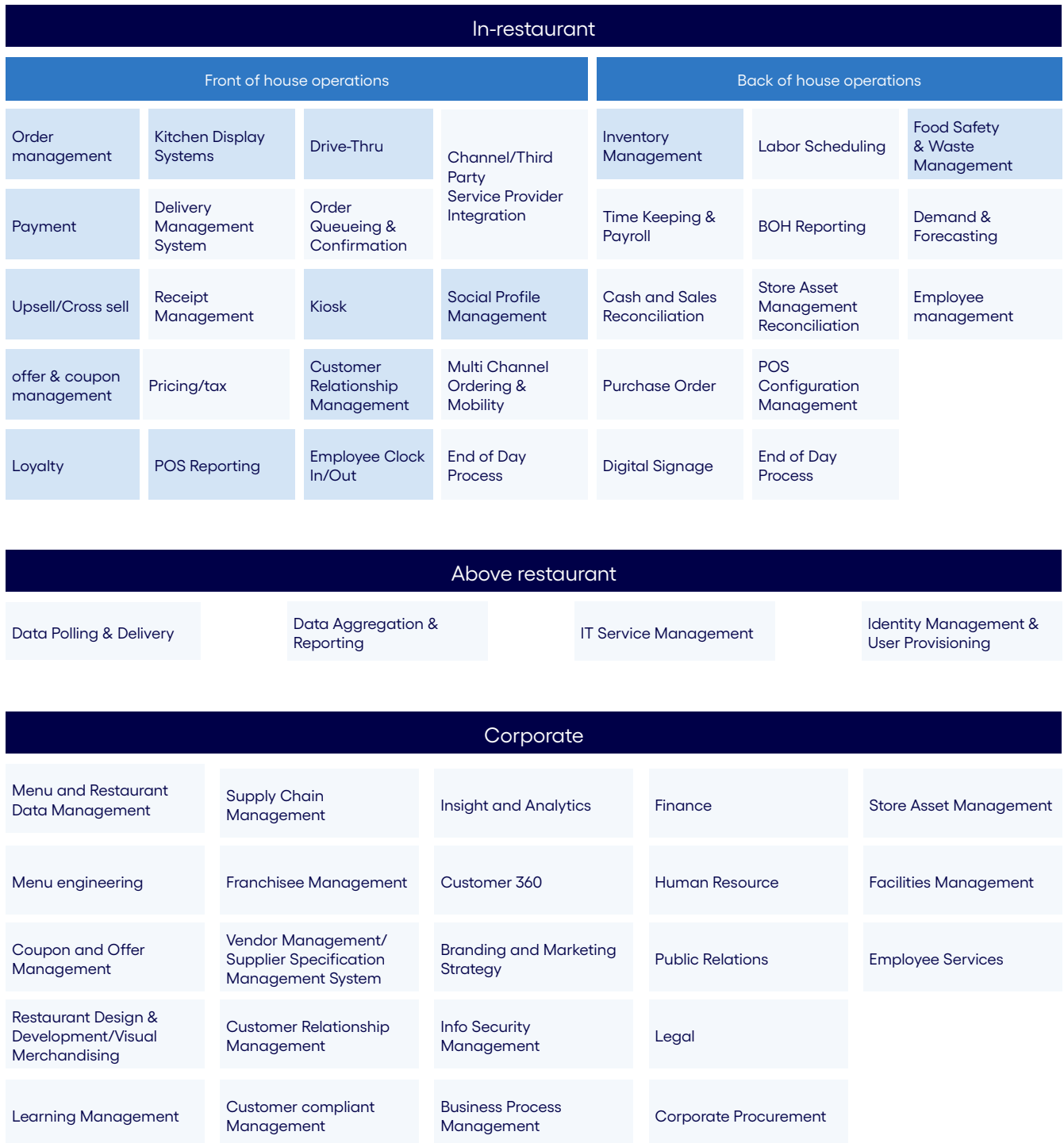


Diagram 3: Restaurant Business Map with GenAI opportunities highlighted.



Reference Architecture

The following are the key components of the reference architecture:

A **GenAI cloud server** that runs the models and applications for tasks like menu generation, order prediction, customer segmentation, etc. The cloud server also keeps and processes the data from the Edge devices and sends them feedback and updates.

A **local network of Edge devices that operate the GenAI models** and applications at the restaurant level, such as kiosks, tablets, cameras, speakers, etc. The Edge devices use compact AI models to do tasks like face recognition, voice recognition, sentiment analysis, etc. The Edge devices also talk to each other and to the cloud server via Wi-Fi or cellular connection.

A set of **sensors and actuators** that gather data from the physical environment, such as temperature, humidity, noise, motion, etc. The sensors and actuators also regulate the physical aspects of the restaurant, such as lighting, heating, ventilation, etc.

The reference architecture can enable the following example use cases:

- A customer walks up to a kiosk and is identified by the face recognition model. The kiosk shows a customized menu created by the Generative AI model based on the customer's preferences, history, and context. The customer orders using voice recognition and pays using biometric authentication.
- A tablet on a table senses a customer and turns on the speaker. The speaker welcomes the customer and offers a suggestion created by the Generative AI model based on the customer's profile, mood, and time of day. The customer can talk with the speaker using natural language and order.
- A camera tracks the crowd size and behavior in the restaurant and sends the data to the Generative AI model. The Generative AI model estimates the demand and supply of food items and changes the inventory and production accordingly. The Generative AI model also improves the staffing and scheduling of the restaurant based on the data.
- A sensor records the temperature and humidity in the kitchen and sends the data to the Generative AI model. The Generative AI model manages the heating and ventilation system to keep the optimal conditions for food preparation and safety. The Generative AI model also warns the staff if any abnormality or hazard is detected.

Technical Architecture

Large Language Models (LLMs) on Edge Devices:

LLMs on Edge Devices can provide more speed, better privacy and security, and online and offline functionality. However, there are some challenges that need to be resolved, including hardware limits, energy use, maintenance, and ethics issues.

Quantization: Quantization is a method to shrink the model size and make it more efficient for use on Edge devices. It uses a technique that lowers the precision of numerical values to lower the computational and memory requirements of AI models. Quantization can be applied at different levels, such as weights, activations, or outputs. Quantization can also be performed at different stages, such as during training, after training, or during inference. Quantization can impact the accuracy, speed, and size of AI models.



Diagram 4: Technical Reference View



The Generative AI cloud server is the central component of the architecture, as it hosts the main models and applications for restaurant management and optimization. The cloud server uses a variety of AI techniques, such as natural language processing, computer vision, machine learning, and Generative AI, to create and improve the solutions for the restaurant. The cloud server also communicates with the Edge devices via APIs or MQTT messages, sending them feedback, updates, and commands.

The Edge devices are the peripheral components of the architecture, as they run the Generative AI models and applications at the restaurant level. The Edge devices use quantized AI models to perform tasks that require low latency, high privacy, or offline availability, such as face recognition, voice recognition, sentiment analysis, etc. The Edge devices also communicate with each other and with the sensors and actuators via Bluetooth, Zigbee, or Wi-Fi, exchanging data and information.

The sensors and actuators are the physical components of the architecture, as they collect and control data from the environment. The sensors and actuators use simple protocols, such as GPIO, I2C, or SPI, to connect with the Edge devices, sending them signals and receiving instructions. The sensors and actuators also enable the GenAI models and applications to interact with the physical aspects of the restaurant, such as lighting, heating, ventilation, etc.



Technology options

Google Coral: This is a platform that offers a range of products, such as a development board, a USB accelerator, and a system-on-module, which can run TensorFlow Lite models at the Edge. It can be used as an Edge device to enable Generative AI capabilities such as face detection, object recognition, and sentiment analysis for QSR kiosks and other restaurant devices. Some advantages of Google Coral are its ease of use, scalability, and integration with Google Cloud services. Some disadvantages are its limited support for other frameworks and languages, its dependency on Google's ecosystem, and its new and evolving status.

NVIDIA Jetson Nano: This is a potent and energy-efficient platform that can run multiple neural networks in parallel and process high-resolution data from multiple sensors. It can be used as an Edge device to boost Generative AI tasks such as computer vision, natural language processing, and speech recognition for QSR kiosks and other restaurant devices. Some advantages of NVIDIA Jetson Nano are its high performance, low power consumption, and compatibility with popular frameworks and tools. Some disadvantages are its higher cost, complexity, and learning curve, as well as its potential overheating and instability issues.

Raspberry Pi: This is a low-cost, small, and adaptable single-board computer that can run Linux-based operating systems and support various programming languages. It can be used as an Edge device to host GenAI models and applications for QSR kiosks and other restaurant devices. Some advantages of Raspberry Pi are its cost-effectiveness, mobility, versatility, and large community support. Some disadvantages are its limited processing power, memory, and storage, as well as its reliance on external peripherals and power sources.

Google Anthos: This is a platform that enables the deployment and management of cloud-native applications across different environments, such as on-premises, public cloud, or Edge devices. It can be used as an Edge device to run Generative AI models and applications for QSR kiosks and other restaurant devices with consistent policies and security. Some advantages of Google Anthos are its portability, scalability, and integration with Google Cloud services. Some disadvantages are its high cost, complexity, and dependency on Google's ecosystem. Google Anthos supports Kubernetes, which is a framework that offers various options for quantization, such as operator-level quantization, model-level quantization, and graph-level quantization.

The technology choices described above have different implications for quantization:

Google Coral: This device has a dedicated TPU, which means that it can run TensorFlow Lite models at the Edge with high speed and low latency. Therefore, quantization is required for Google Coral, as it can enable the device to run the models on the TPU. However, quantization can also limit the flexibility and compatibility of Google Coral, as it can restrict the choice of frameworks and languages. Google Coral supports TensorFlow Lite, which is a framework that offers various options for quantization, such as full integer quantization, float fallback quantization, and hybrid quantization.

NVIDIA Jetson Nano: This device has a powerful GPU, which means that it can run high-resolution and parallel AI models effectively. Therefore, quantization may not be needed for NVIDIA Jetson Nano, as it can handle the computational and memory demands of large models. However, quantization can still be beneficial for NVIDIA Jetson Nano, as it can reduce the power consumption and increase the battery life of the device. NVIDIA Jetson Nano supports TensorFlow, PyTorch, and ONNX, which are frameworks that offer various options for quantization, such as quantization-aware training, quantization emulation, and quantization export.

Raspberry Pi: This device has a limited CPU and GPU, which means that it cannot run complicated or large AI models efficiently. Therefore, quantization can be helpful for Raspberry Pi, as it can reduce the model size and improve the inference speed. However, quantization can also cause accuracy loss, which can affect the performance of Generative AI tasks. Raspberry Pi supports TensorFlow Lite, which is a framework that offers various options for quantization, such as post-training quantization, dynamic range quantization, and integer-only quantization.

Google Anthos: This is a platform that allows the deployment and management of AI applications across different cloud providers and on-premise environments. Therefore, quantization can be useful for Google Anthos, as it can enable the portability and scalability of AI models across heterogeneous hardware and software platforms. However, quantization can also introduce some challenges for Google Anthos, such as ensuring the consistency and compatibility of quantized models across different frameworks and languages. Google Anthos supports TensorFlow, PyTorch, and Scikit-learn, which are frameworks that offer various options for quantization, such as mixed precision training, quantization-aware fine-tuning, and model optimization tools.



Application of Edge and Generative AI use cases in types of restaurants

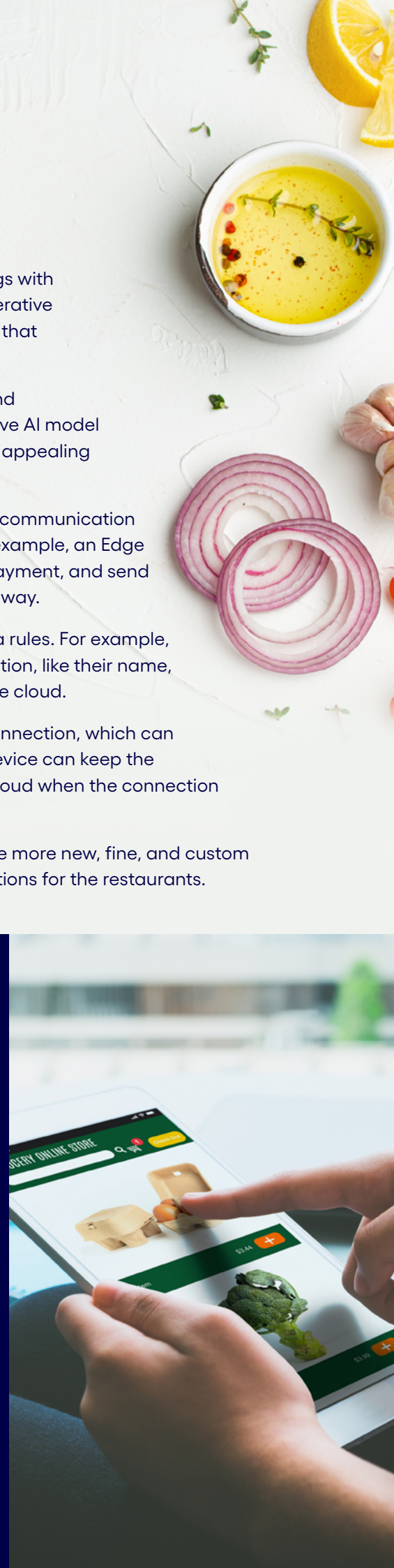
Fine dining

- Generative AI can make new and different recipes, menus, and pairings with the ingredients, cuisines, seasons, and occasions. For example, a Generative AI model can create a dish with unusual tastes and textures, or a wine that matches a dessert.
- Generative AI can also make the food look better by making artistic and attractive designs, colors, and arrangements. For example, a Generative AI model can use edible flowers, sauces, and garnishes to make a more visually appealing effect.
- Edge Computing can speed up and improve the data processing and communication between the restaurant's front-end and back end, and the cloud. For example, an Edge device can handle the customer's reservation, order, feedback, and payment, and send them to the kitchen, the management, and the loyalty program right away.
- Edge Computing can also protect the customer's data and follow data rules. For example, an Edge device can hide and change the customer's personal information, like their name, email, phone number, and payment details, before sending them to the cloud.
- Edge Computing can also let the restaurant work offline or with low connection, which can make the service more available and reliable. For example, an Edge device can keep the important data and functions locally and synchronize them with the cloud when the connection is back.

With Generative AI and Edge Computing, fine dining restaurants can give more new, fine, and custom experiences for the customers, and faster, correct, and lower-cost operations for the restaurants.

Quick serve restaurant

- Using computer vision and natural language processing to identify the customer's face, voice, and order, and suggest customized recommendations and offers.
- Using machine learning and reinforcement learning to change the menu, the prices, and the promotions based on the demand, the season, and the competition.
- Using sensors and actuators to check and manage the temperature, the humidity, and the hygiene of the food and the equipment, and to notify the staff of any issues or anomalies.
- Using chatbots and virtual assistants to help the customers and the employees with their questions, complaints, and suggestions, and provide feedback and guidance.
- Using data analytics and dashboarding to measure and show the performance, the trends, and the outcomes of the restaurant, and spot areas for improvement and innovation.





Application of Edge and GenAI Use Cases in key restaurant functions

QSR kiosks

- Generative AI can create tailor-made menus, deals, and suggestions based on the customer's preferences, behavior, location, and time of the day. For instance, a Generative AI model can recommend a low-calorie salad for a customer who cares about their health or a combo meal for a family with children.
- Generative AI can also improve the user interface and interaction of the kiosks by creating natural language responses, voice synthesis, facial expressions, and gestures. For example, a Generative AI model can welcome the customer, take their order, verify their payment, and express their gratitude for their visit.
- Edge Computing can enable quicker and more dependable data processing and communication between the kiosks and the kitchen, as well as the cloud. For instance, an Edge device can process the customer's order, transmit it to the kitchen, and update the inventory and sales data in real-time.
- Edge Computing can also provide more privacy and security for the customer's data, as well as compliance with data regulations. For instance, an Edge device can encrypt and anonymize the customer's personal information, such as their name, email, phone number, and payment details, before sending it to the cloud.
- Edge Computing can also allow the kiosks to work offline or in low connectivity scenarios, which can enhance the availability and resilience of the service. For instance, an Edge device can store the essential data and functions locally and synchronize them with the cloud when the connection is restored.

By using Generative AI and Edge Computing, QSR kiosks can offer more personalized, interactive, and convenient experiences for the customers, as well as more efficient, precise, and cost-effective operations for the restaurants.

Restaurant POS

The point-of-sale (POS) system is a key element of any restaurant, as it handles the payments, orders, inventory, and customer data. However, traditional POS systems are often old-fashioned, slow, and prone to mistakes and breaches. By using Edge Computing and Generative AI, restaurants can upgrade their POS systems into smart, fast, and secure platforms that can improve the customer experience and business efficiency.

Quicker and more dependable data processing and communication: Edge Computing can solve the latency and bandwidth problems that often impact cloud-based POS systems, especially during busy times or network outages. By processing the data locally on the Edge devices, the POS system can work quicker and more dependably, ensuring smooth payments and orders.

Better privacy and security of customer data: Edge Computing can also safeguard customer data from unauthorized access or leakage, as it reduces the exposure of sensitive information to the cloud or the internet. By encrypting and anonymizing the data on the Edge devices, the POS system can follow data regulations and avoid identity theft, fraud, or cyberattacks.

Improved customization and interaction of customer service: Generative AI can enable the POS system to provide more personalized and engaging service for customers, by using natural language processing, computer vision, and speech recognition to understand and respond to customer needs and preferences. For example, a Generative AI model can welcome the customer by their name, offer them relevant discounts or loyalty rewards, recommend items based on their order history or dietary limitations, and generate natural and human-like conversations.

Enhanced efficiency and accuracy of restaurant operations: Generative AI can also help the POS system optimize restaurant operations, by using data analytics, machine learning, and reinforcement learning to monitor and improve the performance, trends, and outcomes. For example, a Generative AI model can track and manage the inventory, supply chain, and waste, predict the demand, adjust the prices and promotions, and provide feedback and suggestions for the staff.





One example of how Generative AI, AI and Edge Computing can help improve restaurant POS is as follows:

- A customer walks into a fast-food restaurant and scans a QR code on the table with their smartphone. The QR code directs them to a web app that serves as a POS system for the restaurant. The web app runs on the Edge device, which is a small server located in the restaurant. The Edge device processes the data locally and communicates with the cloud only when necessary, ensuring fast and reliable service.
- The web app greets the customer by their name and shows them a menu that is customized based on their previous orders, preferences, and allergies. The customer can use voice or text to place their order, and the web app uses natural language processing and speech recognition to understand and confirm their order. The web app also offers them a discount coupon that they can apply to their order and suggests some upsell items that they might like.
- The customer pays for their order using their credit card or digital wallet, and the web app securely encrypts and verifies their payment information on the Edge device, without sending it to the cloud or the internet. The web app also updates the customer's loyalty points and rewards, and thanks them for their purchase.
- The web app sends the order details to the kitchen staff, who prepare the food and deliver it to the customer's table. The web app also monitors the inventory and supply chain of the restaurant and alerts the manager if any item is running low or needs to be reordered. The web app also analyzes the sales data and customer feedback and provides insights and recommendations to the manager on how to optimize the menu, pricing, and promotions.
- The customer enjoys their meal and leaves the restaurant satisfied. The web app asks them to rate their experience and provide any suggestions or complaints. The web app uses natural language generation and sentiment analysis to generate a friendly and personalized response to the customer, and to address any issues or concerns they might have. The web app also learns from the customer's feedback and preferences and improves its service for the next time.



Restaurant contact centers

Edge Computing combined with Generative AI can enhance the efficiency of restaurant contact centers that handle customer inquiries, feedback, and problems. Through Edge Computing, data processing becomes faster, less costly, more secure, and accessible even without an internet connection. Generative AI can use natural language, speech, and vision to give more accurate, personal, and friendly responses to customers, and to handle repetitive tasks. Generative AI can also use data to understand and satisfy customers better, and to suggest suitable offers and solutions.

Examples of how Generative AI, AI and Edge Computing can help improve restaurant contact centers include:

- A customer who has ordered food online from a restaurant may receive a phone call or a text message from the contact center, which uses Generative AI to start a natural and friendly conversation.
- The contact center can use speech recognition and natural language understanding to recognize the customer's voice, name, and order details, and to confirm that the order has been delivered on time and as expected.
- The contact center can also use sentiment analysis to detect the customer's mood and satisfaction level, and to offer appropriate apologies, compliments, or incentives. The contact center can use Edge Computing to run the Generative AI models locally on the device, without relying on the internet connection or the cloud server, which can reduce latency, cost, and security risks.
- The contact center can also use Edge Computing to store and process the customer's data and feedback, and to update and improve the Generative AI models over time.

By using Generative AI, AI and Edge Computing, the restaurant contact center can enhance the customer experience, increase customer loyalty, and reduce the human workload.

Conclusion

Edge Computing, AI, and Generative AI are revolutionizing the future of restaurant technology by providing solutions to the challenges faced by the industry. By bringing computing power closer to where data is generated, Edge Computing reduces latency and enables faster decision-making. In-restaurant cloud technology, driven by Edge Computing, can help restaurants process data faster and more reliably, enhancing system uptime. Large Language Models (LLMs) on Edge devices provide more speed, better privacy and security, and online and offline functionality. However, there are challenges that need to be resolved, including hardware limits, energy use, update and maintenance issues, and ethics issues. Quantization is a method to shrink the model size and make it more efficient for use on Edge devices. Overall, the integration of Edge Computing, AI, and Generative AI in the restaurant industry has the potential to enhance profitability by reducing costs and improving processes.

Glossary

LLM	Large Language Model
GenAI	Generative AI
API	Application Programming Interface
QSR	Quick Serve Restaurant
POS	Point of Sale
TPU	Tensor Processing Unit
CPU	Central Processing Unit
GPU	Graphics Processing Unit
IoT	Internet of Things
MQTT	Message Queuing Telemetry Transport
GPIO	General Purpose Input / Output
I2C	Inter Integrated Circuit
SPI	Serial Peripheral Interface

References

1. <https://coral.ai/>
2. <https://www.tensorflow.org/lite>
3. <https://www.nvidia.com/en-in/autonomous-machines/embedded-systems/jetson-nano/product-development/>
4. <https://www.raspberrypi.com/news/iot-gets-a-machine-learning-boost-from-Edge-to-cloud/>
5. https://huggingface.co/docs/optimum/en/concept_guides/quantization
6. <https://www.grandviewresearch.com/press-release/global-Edge-computing-market>
7. <https://insights.refed.org/uploads/documents/refed-insights-engine-food-waste-monitor-methodology-vfinal-2022-03-11.pdf>
8. <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consumer-business/us-cb-restaurant-loyalty-program.pdf>
9. <https://restaurant.org/research-and-media/research/research-reports/state-of-the-industry/>
10. <https://cloud.google.com/anthos>

Author



Aditya has over 24 years of experience working as an Enterprise / Solution Architect for Retail, Consumer Goods and T&H industry. His primary focus area is around modernization of Restaurant Technology for both QSRs and Fine Dining Restaurants. He has keen interest in the evolution of AI and GenAI technologies and has been involved in extensive research work on how to leverage these new technologies for improved Restaurant operations and customer experience.

Aditya Sankar Sarkar

(Principal Architect, Travel & Hospitality)



Cognizant (Nasdaq:100; CTSI) engineers modern businesses. We help our clients modernize technology, reimagine processes and transform experiences so they can stay ahead in our fast-changing world. Together, we're improving everyday life. See how at www.cognizant.com or [@cognizant](https://twitter.com/cognizant).

World Headquarters

300 Frank W. Burr Blvd.
Suite 36, 6th Floor
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277

European Headquarters

280 Bishopsgate
London
EC2M 4RB, England
Tel: +44 (0) 20 7297 7600

India Operations Headquarters

5/535, Okkiam Thoraiyakkam,
Old Mahabalipuram Road,
Chennai 600 096
Tel: 1-800-208-6999
Fax: +91 (0) 44 4209 6060

APAC Headquarters

1 Fusionopolis Link,
Level 5 NEXUS@One-North,
North Tower, Singapore 138542
Phone: +65 6812 4000

© Copyright 2024, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission of Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned here in are the property of their respective owners.